Normalization and Subtraction: Two Approaches to Facilitate Gene Discovery

Maria de Fatima Bonaldo,[1] Gregory Lennon,[3] and Marcelo Bento Soares[1,2,4]

[1]Department of Psychiatry, College of Physicians and Surgeons of Columbia University, and [2]The New York State Psychiatric Institute, New York, New York 10032; [3]Human Genome Center, Lawrence Livermore National Laboratory, Livermore, California 94551

Large-scale sequencing of cDNAs randomly picked from libraries has proven to be a very powerful approach to discover (putatively) expressed sequences that, in turn, once mapped, may greatly expedite the process involved in the identification and cloning of human disease genes. However, the integrity of the data and the pace at which novel sequences can be identified depends to a great extent on the cDNA libraries that are used. Because altogether, in a typical cell, the mRNAs of the prevalent and intermediate frequency classes comprise as much as 50–65% of the total mRNA mass, but represent no more than 1000–2000 different mRNAs, redundant identification of mRNAs of these two frequency classes is destined to become overwhelming relatively early in any such random gene discovery programs, thus seriously compromising their cost-effectiveness. With the goal of facilitating such efforts, previously we developed a method to construct directionally cloned normalized cDNA libraries and applied it to generate infant brain (INIB) and fetal liver/spleen (INFLS) libraries, from which a total of 45,192 and 86,088 expressed sequence tags, respectively, have been derived. While improving the representation of the longest cDNAs in our libraries, we developed three additional methods to normalize cDNA libraries and generated over 35 libraries, most of which have been contributed to our Integrated Molecular Analysis of Genomes and Their Expression (IMAGE) Consortium and thus distributed widely and used for sequencing and mapping. In an attempt to facilitate the process of gene discovery further, we have also developed a subtractive hybridization approach designed specifically to eliminate (or reduce significantly the representation of) large pools of arrayed and (mostly) sequenced clones from normalized libraries yet to be (or just partly) surveyed. Here we present a detailed description and a comparative analysis of four methods that we developed and used to generate normalize cDNA libraries from human (15), mouse (3), rat (2), as well as the parasite *Schistosoma mansoni* (1). In addition, we describe the construction and preliminary characterization of a subtracted liver/spleen library (INFLS-S1) that resulted from the elimination (or reduction of representation) of ~5000 INFLS–IMAGE clones from the INFLS library.

Large-scale single-pass sequencing of cDNA clones randomly picked from libraries has proven to be a powerful approach to discover genes (Adams et al. 1991, 1993a,b, 1995; Khan et al. 1992; McCombie et al. 1992; Okubo et al. 1992; Matsubara and Okubo 1993; see also Hillier et al., this issue). However, the significance of using cDNA libraries that are well suited for this purpose should not be underestimated (Adams et al. 1993b).

Ordinary cDNA libraries may contain a high frequency of undesirable ("junky") clones (Adams et al. 1991, 1992) that may not only drastically impair the overall efficiency of the approach, but also seriously compromise the integrity of the data that are generated. Among such junky clones are: (1) clones that consist exclusively of poly(A) tails of mRNAs; (2) clones that contain very short cDNA inserts; (3) clones that contain nothing but the 3′ half of the *Not*I–oligo$(dT)_{18}$ primer used for synthesis of first-strand cDNA ligated to an adaptor; and (4) chimeric clones, i.e., cDNAs derived from different mRNAs joined artifactually during ligation . Furthermore, given that, as a general rule, the frequency of occurrence of a cDNA clone in a library is equivalent to that of its corresponding mRNA in the cell, even high-quality cDNA libraries may not be ideal for large-scale sequencing.

[4]Corresponding author.
E-MAIL cuc cuccfa.ccc.columbia.edu; FAX (212) 781-3577.

6:791–806 1996 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/96 $5.00

Reassociation-kinetics analysis indicates that the mRNAs of a typical somatic cell are distributed in three frequency classes: (1) superprevalent (consisting of about 10–15 mRNAs that altogether represent 10–20% of the total mRNA mass); (2) intermediate (1000–2000 mRNAs; 40–45%); and (3) complex (15,000–20,000 mRNAs; 40–45%) (Bishop et al. 1974; Davidson and Britten 1979). Accordingly, once most mRNAs of the prevalent and intermediate frequency classes are identified, redundancy levels are expected to become greater than 60%. For this reason, the use of normalized libraries, in which the frequency of all clones is within a narrow range (Soares et al. 1994), has been shown to be beneficial for large-scale sequencing (Berry et al. 1995; Houlgatte et al. 1995). Calculations show that at $C_0t = 5.5$ [where C_0 is the total DNA concentration and t is the time (moles nucleotides per liter × sec)], of the three kinetic classes of mRNAs, the most abundant species are diminished drastically, while all frequencies are brought within the range of one order of magnitude (Soares et al. 1994).

However, because a large fraction of all human genes has been identified already, redundant identification of genes that are expressed in multiple tissues cannot be avoided simply by the use of normalized libraries. Hence, we argue that the use of subtractive cDNA libraries enriched for genes expressed at low levels and that have not yet been identified should become increasingly more advantageous for large-scale sequencing programs.

While attempting to improve the representation of the longest cDNAs in our libraries, we developed three methods for construction of normalized libraries, in addition to the procedure that we described previously (Soares et al. 1994), and used them successfully to generate normalized cDNA libraries from human (15), mouse (3), rat (2), and *Schistosoma mansoni* (1) tissues. All human and mouse cDNA libraries have been contributed to the Integrated Molecular Analysis of Genomes and Their Expression (IMAGE) Consortium (Lennon et al. 1996), and to date a total of 315,408 expressed sequence tags (ESTs) have been derived from these libraries (dbEST release 052396; http://www.ncbi.nlm.nih.gov).

Here we present a detailed description and a comparative analysis of the four methods that we have developed to normalize cDNA libraries; we describe a simple procedure for the construction of subtractive cDNA libraries; and we discuss strategies that take advantage of subtractive hybridization to expedite the ongoing IMAGE/Washington University/Merck gene discovery program.

RESULTS

While attempting to improve the representation of the longest cDNAs in our normalized libraries, we developed four methods and constructed over 35 libraries, most of which are described here. A list comprising 15 human, three mouse, two rat, and one schistosome library with their respective names, number of recombinants, sequence tags, and methods used for normalization and preparation of single-stranded plasmids is shown in Table 1.

Extensive characterization of two normalized libraries [normalized infant brain (1NIB) and normalized fetal spleen (1NFLS)] constructed according to our previously described procedure (Soares et al. 1994; here designated as method 1) confirmed our original observations that a great extent of normalization can be achieved with this method for most cDNA species (e.g., cf. lanes 9,10 in Fig. 1M–P). It is noteworthy that the frequency of cDNA 122 (used as the probe in P) was increased with normalization from <0.0006% in the starting library to 0.007% in the 1NIB library (Soares et al. 1994). However, Southern hybridization of starting and normalized libraries with a battery of cDNA probes revealed that on occasion truncated clones were favored over their longest counterparts during the process. This was first observed when Southern blots of *Not*I + *Hin*dIII-digested plasmid DNA from starting and normalized infant brain libraries were hybridized with a cDNA probe for mitochondrial 16S rRNA (see Fig 1L, lanes 9,10). Not only was the frequency of these mitochondrial cDNA clones not reduced effectively during the process of normalization (frequency of occurrence in starting and normalized infant brain libraries was 1.4% and 1.0%, respectively), but also the length of the hybridizing cDNAs was noticeably smaller in the normalized library. Comparative sequence analysis (not shown) of a number of hybridizing mitochondrial 16S rRNA clones from both starting and normalized libraries revealed that whereas the 3′ end of most cDNAs derived from the starting library corresponded to the bona fide 3′ end of the 16S rRNA, the 3′ end of the majority of the cDNAs isolated from the normalized library corresponded to sequences further upstream on the

Table 1. Complete List and Main Features of the Normalized Human, Mouse, Rat, and Schistosome cDNA Libraries

mRNA source	Normalized library name	Number of recombinants in the normalized library	Preparation of single-stranded plasmids	Method of normalization	Library tag[a]
Human infant brain[b]	1NIB	2,500,000	in vivo	1	AGGAA
Human fetal liver spleen[c]	Nb2HFLS20W (1NFLS)	19,000,000	in vivo	1	AGATCT
	5Nb2HFLS20W	3,200,000	in vitro	2-1	
	6Nb2HFLS20W	1,400,000	in vitro	2-3	
	14Nb2HFLS20W	3,200,000	in vitro	4	
	15Nb2HFLS20W	35,000,000	in vitro	2-2	
Human term placenta	Nb2HP	750,000	in vivo	2-1	AGGAA
Human 8-9W placenta	2NbHP8-9W	100,000	in vitro	2-3	GA
Human breast[d]	2NbHbst-3NbHBst[e]	2,090,000	in vivo	2-1	CC
Human adult brain[f]	N2b4HB55Y-N2b5HB55Y[g]	3,170,000	in vivo	2-1	GC
Human retina[h]	2N2b4HR-N2b5HR	1,600,000	in vivo	2-1	AC
Human pineal gland[i]	3NbHPG	1,000,000	in vitro	2-1	CG
Human ovary tumor[j]	NbHOT	1,100,000	in vivo	2-1	GG
Human melanocytes[k]	2NbHM	6,800,000	in vitro	2-3	AG
Human fetal heart[l]	NbHH19W	9,700,000	in vitro	4	ATC
Human parathyroid adenoma[m]	NbHPA	3,400,000	in vitro	4	ACCAA
Human senescent figroblast[n]	NbHSF	9,900,000	in vitro	4	AACCA
Human multiple sclerosis plaques[o]	2NbHMSP	1,100,000	in vitro	3	CA
Human fetal lung[l]	NbHL19W	21,700,000	in vitro	4	AA
19.5-dpc mouse embryo[p]	p3NMF19.5	3,400,000	in vitro	4	ACAAC
17.5-dpc mouse embryo[p]	NbME17.5	6,800,000	in vitro	4	GACAC
13.5- to 14.5-dpc mouse embryos[p]	NbME13.5-14.5	380,000	in vitro	4	GGAAA
Rat heart[q]	NbRH	400,000	in vitro	4	ACAAC
Rat kidney[q]	2NbRK	130,000	in vitro	4	CAAAC
8-week-old adult schistosome[r]	NbS8W	1,000,000	in vitro	4	GAAAG

With the exception of 1NIB, which was constructed in the Lafmid BA vector, all libraries were constructed in the pT7T3–*Pac* vector. Cloning sites were *Not*I and *Eco*RI, except for fetal liver spleen (*Pac*I and *Eco*RI) and infant brain (*Not*I and *Hind*III).

[a]The library tag is a sequence identifier present in the oligonucleotide used to prime the synthesis of first-strand cDNA, between the recognition sequence for the rare restriction enzyme (*Not*I or *Pac*I in the case of the liver spleen library) used for directional cloning and the dT_{18} stretch (or dT_{25} in the human parathyroid adenoma, senescent fibroblast, mouse embryo, rat, and *Schistosoma mansoni* libraries) located at the 3′ end of the primer.

[b]Human infant brain (kindly provided by Dr. Conrad Gilliam, Columbia University, New York, NY) was from a 72-day-old female who died in consequence of spinal muscular atrophy.

[c]Human fetal liver spleen (kindly provided by Dr. Stephen Brown, Columbia Presbyterian Medical Center, New York, NY) was from a 20-week-old postconception normal female.

[d]Total cellular poly(A)$^+$ mRNA from normal breast pooled from reduction mammoplasty tissue was kindly provided by Dr. Anne Bowcock and Ms. Monique Spillman, University of Texas Southwestern Medical Center at Dallas.

[e]2NbHbst differs from 3HbHbst in the C_0t used for hybridization (237 and 20, respectively).

[f]Total cellular adult brain RNA (kindly provided by Dr. Donald Gilden, University of Colorado Health Sciences Center, Denver) was obtained from a 55-year-old male who died of a ruptured aortic aneurysm. Brain tissue (frontal, parietal, temporal, and occipital cortex from the left and right hemispheres, subcortical white matter, basal ganglia, thalamus, cerebellum, midbrain, pons, and medulla) was acquired 17-18 hr after death.

[g]N2b4HB55Y and 2N2b4HR differ from N2b5HB55Y and N2b5HR, respectively, in the average size of their cDNA inserts (1.5–2.5 kb and 0.4–1.5 kb, respectively).

[h]Total cellular normal human retina RNA (kindly provided by Dr. Roderick R. McInnes, University of Toronto and Hospital for Sick Children, Canada) was obtained from a 55-year-old Caucasian male.

[i]Human pineal gland [kindly provided by Dr. David Klein, National Institute of Child Health and Human Development, National Institutes of Health (NIH)] was derived from a group of three pineal glands (gland 1: 48-year-old Caucasian male: gland 2: 18-year-old Caucasian female, gland 3: 20-year-old African American male).

[j]Total cellular human ovary tumor mRNA was kindly provided by Dr. Anne Bowcock and Ms. Monique Spillman, University of Texas Southwestern Medical School. It was obtained from a 36-year-old Caucasian with a papillary serous cystadenocarcinoma grade III with surface extensions and metastases.

[k]Total cellular human melanocyte RNA (kindly provided by Dr. Anthony Albino and Dr. Alice de Oliveira, Memorial Sloan-Kettering Cancer Center, New York, NY) was derived from normal foreskin.

[l]Normal human fetal heart and lung (kindly provided by Dr. Stephen Brown, Columbia Presbyterian Medical Center) were derived from the same 19-week-postconception speciman.

[m]Human parathyroid tumor (kindly provided by Dr. Stephen Marx, National Institute of Diabetes and Digestive and Kidney Diseases, NIH) was derived from sporadic adenomas.

[n]Cytoplasmic mRNA from senescent normal human fibroblasts was kindly provided by Dr. Barbara Burhart (National Institute of Environmental Health Sciences, NIH). The cells were prepared by passaging normal human fibroblasts derived from foreskin until they exhibited an enlarged, flattened phenotype and failure to divide (labeling index of <2% following 48 hr-bromodeoxyuridine incorporation).

[o]Total cellular RNA from multiple sclerosis plaques (kindly provided by Dr. Kevin G. Becker, NINDS, NIH) was extracted from four lesions obtained from one patient.

[p]Total cellular RNA for construction of the mouse (C57BL/6J strain) embryonic libraries was kindly provided by Dr. Minoru Ko (Wayne State University, Detroit, MI).

[q]Rat tissues were obtained from an adult Zivic-Miller Sprague Dawley female and were kindly provided by Dr. Stephen Brown (Columbia Presbyterian Medical Center).

[r]Total cellular RNA from mature 8-week-old *Schistosoma mansoni* worms was kindly provided by Dr. Ron Blanton, Case Western Reserve University, Cleveland, OH.

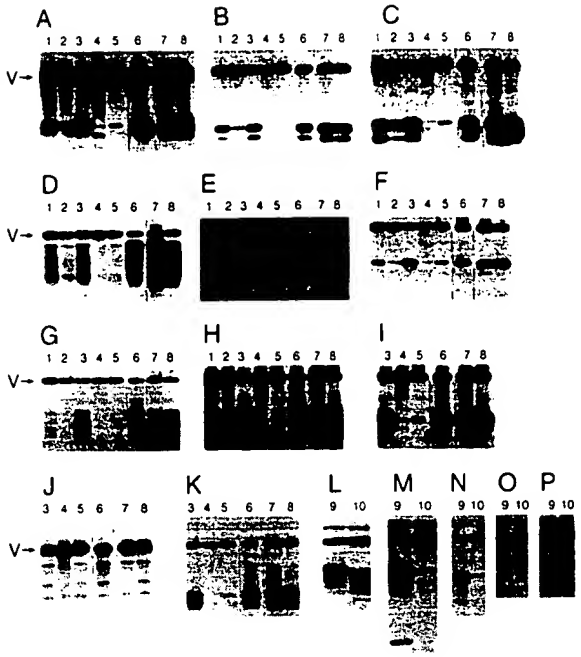

Figure 1 Comparative analysis of starting and normalized cDNA libraries by Southern hybridization with 14 cDNA probes. The 0.015 μg *Pacl* + *EcoRI* digested plasmid DNA from the starting fetal liver/spleen library (lane *6*), from the normalized fetal liver/spleen libraries constructed according to method 2-1 (lane *1*), method 2-3 (lane *2*), method 2-2 (lane *3*), method 1 (lane *4*), method 4 (lane *5*), and from the liver/spleen mini-libraries enriched for abundant cDNAs (HAP-bound fractions) generated with method 2-1 (lane *7*) and method 4 (lane *8*) were electrophoresed on 1% agarose gels, transferred to nylon membranes (GeneScreenPlus; DuPont/NEN) and hybridized at 42°C in 50% formamide, 5× Denhardt's solution, 0.75 M NaCl, 0.15 M Tris (pH 7.5), 0.1 M sodium phosphate, 0.1% sodium pyrophosphate, 2% SDS containing sheared and denatured salmon sperm DNA at 100 μg/ml. Similarly, 0.05 μg *NotI* + *Hind*III digested plasmid DNA from the starting (IB; lane *9*) and normalized (1NIB; lane *10*; method 1) infant brain libraries (Soares et al. 1994) were electrophoresed, transferred, and hybridized as described above. Radioactive probes were prepared by random primed synthesis using the Prime-it II kit (Stratagene). The following probes were used: α-globin (*A*), β-globin (*B*), γ-globin (*C*), serum albumin (*D*, shorter exposure; *E*, longer exposure), acidic ribosomal phosphoprotein PO (*F*), H19 RNA (*G*, shorter exposure; *H*, longer exposure), apolipoprotein A (*I*), angiotensinogen (*J*), unknown cDNA 8 (*K*), mitochondrial 16S rRNA (*L*), α-Tubulin (*M*), myelin basic protein (*N*), secretogranin (*O*), and unknown cDNA 122 (*P*). All probes were contaminated intentionally with a small amount of vector DNA to enable visualization of vector bands and thus confirm that a similar amount of library DNA was loaded in all lanes. (V) vector band, which is released from the cDNA inserts by double digestion with the restriction enzymes specified above.

16S rRNA. The occurrence of such 3′ truncations was also documented by sequence analysis (not shown) for serum albumin cDNAs in the fetal liver/spleen library (see Fig. 1D,E, lanes 4,6).

Reasoning that this problem could be circumvented if the fragments used in the hybridization with the single-stranded circles (1) were in excess, and (2) spanned the entire length of the cDNAs, we developed an alternative procedure to normalize cDNA libraries based on hybridization of in vitro synthesized RNA (driver) from an entire library with the library itself in the form of single-stranded circles (tracer) (see methods 2-1 and 2-2 in Fig. 2). Several normalized libraries were generated by this procedure (see Table 1).

Southern hybridization of endonuclease-restricted plasmid DNA from starting and normalized libraries with a number of cDNA probes (Fig. 1) indicated clearly that these methods effectively improved the representation of the longest cDNAs in the normalized libraries (e.g., cf. lanes 1,4 in Fig. 1A,D,E,G,H). However, characterization of one of these libraries (5Nb2HFLS20W) by colony hybridization with cDNA probes (not shown) indicated that this approach was effective to reduce the frequency of some, but not all, of the most abundant clones (e.g., serum albumin was reduced about 20-fold, whereas γ-globin was reduced only twofold). No difference was observed when hybridizations were performed at different conditions [0.4 M NaCl and 50% formamide at 42°C as in methods 2-1 and 2-3; 0.12 M NaCl, 50% formamide, and 1% sodium dodecyl sulfate (SDS) at 30°C as in method 2-2 (see lane 3 in Fig. 1); 0.4 M NaCl and 80% formamide at 42°C, not shown].

It is noteworthy that Northern hybridization (not shown) of in vitro transcribed RNA synthesized from an entire plasmid library with probes derived from the abundant cDNAs that failed to be normalized effectively by this procedure (e.g., globins in the fetal liver/spleen library and glyceraldehyde-3-phosphate dehydrogenase (G3PD) in the breast library) indicated that they were not as prevalent in the population of in vitro transcribed RNAs as they were in their respective starting cDNA libraries.

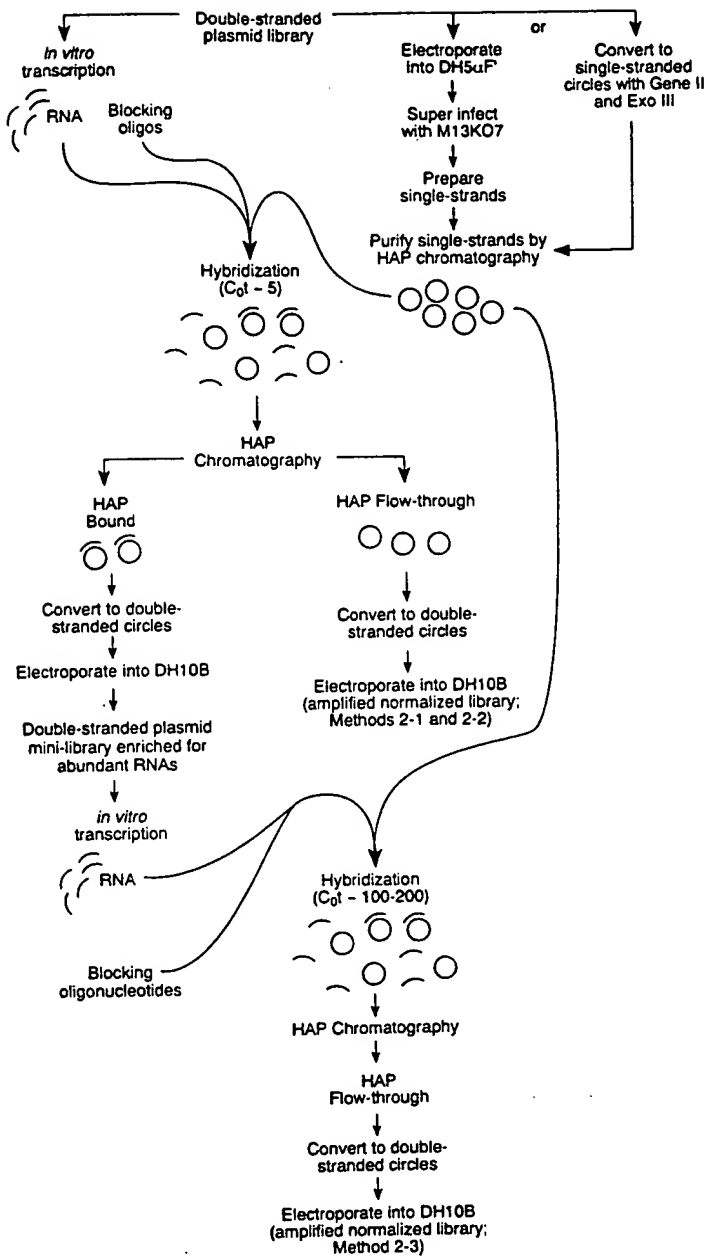

Figure 2 Diagram of the normalization methods 2-1, 2-2, and 2-3. Double-stranded plasmid DNA representing an entire starting library is (1) linearized with either *Sfi*I, *Not*I, or *Pac*I and used as template for synthesis of RNA in vitro using T3 or T7 RNA polymerases, and (2) converted to single-stranded circles either in vivo, upon electroporation into DH5αF' and superinfection with M13KO7, or in vitro by the combined action of Gene II and Exonuclease III (Life Technologies). Single-stranded plasmid DNA is HAP-purified and hybridized (C_0t ~ 5) with excess RNA (pretreated with RNase-free DNAse I; Promega), blocked with appropriate oligonucleotides to prevent hybridization through common vector sequences (see Methods section). Both the fraction that remains single-stranded (flow-through) as well as the resulting hybrids (bound) are purified by HAP chromatography. The HAP flow-through fraction is converted to double-stranded plasmids, electroporated into DH10B bacteria (Life Technologies), and propagated under ampicillin selection to generate an amplified normalized library (methods 2-1 and 2-2, depending on the conditions used for hybridization; see Methods section). The HAP-bound fraction is also converted similarly to double-stranded plasmids, electroporated into bacteria, and propagated under ampicillin selection to generate a mini-library enriched for abundant cDNAs. Double-stranded plasmid DNA from this mini-library is linearized and used as template for synthesis of RNA in vitro. After digestion of the plasmid DNA template with ribonuclease-free DNAse I (Promega), the RNA (driver) is blocked with appropriate oligonucleotides and hybridized (C_0t ~100–200) with HAP-purified single-stranded plasmids derived from the starting library (see above). The remaining single-stranded circles are purified by HAP chromatography, converted to double-stranded circles, electroporated into DH10B (Life Technologies), and propagated under ampicillin selection to generate an amplified normalized library (method 2-3).

A significantly improved extent of normalization was achieved when runoff RNA synthesized from the plasmid mini-library enriched for abundant cDNAs (hydroxyapatite (HAP)-bound fraction of method 2-1 in Fig. 2) was hybridized (C_0t = 100–200) with single-stranded circles from the starting library (see method 2-3 in Fig. 2 and Table 1; cf. lanes 1,2 in Fig. 1A–D,F,G).

In an effort to preserve the positive characteristics of both methods 1 and 2 (i.e., the adequate extent of normalization achieved with method 1, and the improved representation of the longest cDNAs achieved with method 2), we developed two additional reassociation kinetics based procedures involving DNA–DNA hybridization (methods 3 and 4; see Fig. 3).

Method 3, which was successfully used to construct a normalized library from multiple

sclerosis plaques (see 2NbHMSP in Table 1), involved hybridization of a 20-fold excess of single-stranded cDNA fragments (comprising the 5′ halves of all inserts of the starting library, generated by Exonuclease III digestion of gel-purified double-stranded cDNAs; see Fig. 3) with complementary single-stranded circles produced in vitro by the combined action of Gene II and Exonuclease III (Life Technologies).

Southern hybridization of *Not*I + *Eco*RI-digested plasmid DNA from the starting and normalized (with methods 2-1 and 3) multiple sclerosis plaques library with mitochondrial 16S rRNA and myelin basic protein cDNA probes (not shown) clearly indicated that method 3 was superior to method 2-1 in that a much greater extent of normalization was achieved, at the same time that it maintained (similar to method 2-1) appropriate representation of the longest cDNAs in both cases.

For the libraries constructed with method 4 (see Table 1 and Fig. 3), double-stranded cDNA inserts generated by the polymerase chain reaction (PCR) with T3 and T7 primers were melted and hybridized (in the presence of vast excess of blocking oligonucleotides) with single-stranded plasmid library DNA prepared in vitro.

Southern hybridization of *Pac*I + *Eco*RI-digested plasmid DNA from starting and normalized (with methods 1, 2-1–2-3, and 4) fetal liver/spleen libraries (Fig. 1) with several cDNA probes (including those that revealed incomplete normalization with methods 2-1–2-3, such as α-globin, β-globin and γ-globin) demonstrated the efficacy of method 4 in achieving the desired extent of normalization obtained with method 1 (cf. lanes 1–6 in Fig. 1A–D, F–H, and lanes 3–6 in Fig. 1I–K) while preserving the representation of the longest cDNAs (e.g., the longest albumin cDNA was present in the normalized library prepared with method 4, shown in lane 5 of Fig. 1D,E, but it was undetectable in the normalized library constructed with method 1, shown in lane 4; a similarly remarkable difference was revealed with the cDNA probe for H19 RNA, shown in Fig. 1G,H). Characterization of the normalized library generated with method 4 by colony hybridization with 10 cDNA probes (not shown), which occur at a wide range of frequencies in the starting library, confirmed the effectiveness of the procedure to narrow their frequencies down to within one order of magnitude (e.g., the frequencies of the cDNAs for γ-globin, α-globin, β-globin, H19 RNA, and transferrin were reduced

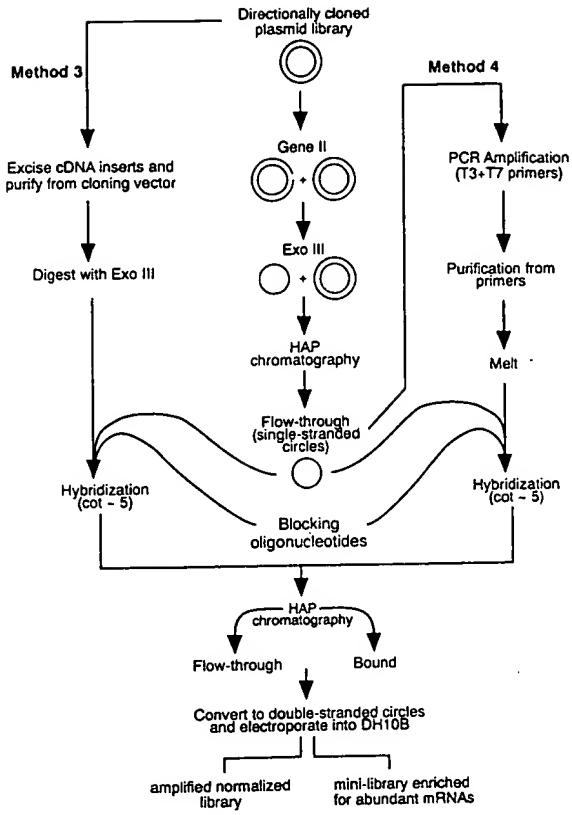

Figure 3 Diagram of the normalization methods 3 and 4. In method 3 double-stranded plasmid DNA from a starting library is digested with restriction enzymes that generate 5′ protruding ends, and the excised cDNA inserts are gel-purified from the cloning vector and digested with Exonuclease III to yield noncomplementary single-stranded fragments, each representing half of a cDNA insert. Note that the single-stranded fragments that span the 5′ half (but not the 3′ half) of the cDNA inserts are complementary to single-stranded plasmids prepared in vitro. These single-stranded DNA fragments are blocked with appropriate oligonucleotides (see Methods) and hybridized with single-stranded library DNA prepared in vitro (*middle* column). The remaining single-stranded circles are HAP-purified, converted to double-stranded plasmids, electroporated into DH10B bacteria (Life Technologies), and propagated under ampicillin selection to generate a normalized library. In method 4, single-stranded library DNA is used as template for PCR amplification with T3 and T7 primers. PCR-amplified cDNAs are purified from excess primers, melted, and hybridized with single-stranded library DNA in the presence of blocking oligonucleotides. The remaining single-stranded circles are purified by HAP chromatography, converted to double-stranded plasmids, electroporated into bacteria, and propagated under ampicillin selection to generate a normalized library.

from 9.2%, 6.4%, 3.6%, 1.8%, and <0.2% to 0.04%, 0.02%, 0.01%, 0.1% and 0.1%, respectively).

In order to assess further the ability of these normalization procedures to preferentially reduce the representation of the most abundant cDNAs, we have performed a comparative sequence analysis (not shown) of 100 clones picked randomly from the fetal liver/spleen cDNA library normalized with method 4 (14Nb2HFLS20W in Table 1; HAP-flow-through fraction in Fig. 3), and from two fetal liver/spleen mini-libraries enriched for abundant cDNAs (HAP-bound fractions in Figs. 2 and 3) obtained during HAP purification of the normalized libraries prepared according to methods 2-1 (5Nb2HFLS20W) and 4 (14Nb2HFLS20W). A number of cDNAs known to be prevalent in the starting fetal liver/spleen library (e.g., albumin, γ-globin, α-globin, β-globin, mitochondrial RNAs, and apolipoproteins A and H) were found at increased frequencies in both mini-libraries enriched for abundant cDNAs, but none of them was represented in the sample of 100 clones from the normalized library. It is noteworthy that while 47% of the sequences derived from the normalized library were not represented in the "all nonredundant" subdivision of sequences of GenBank + EMBL + DDBJ + PDB, the majority of the sequences obtained from the mini-libraries of abundant cDNAs derived from methods 2-1 and 4 (91.4% and 86.9%, respectively) did have homologous sequences in that data base. Furthermore, although 49% of the sequences derived from the normalized library had fewer than 10 homologous ESTs in the dbEST subdivision of GenBank, most of the sequences obtained from both mini-libraries had greater than 10 homologous ESTs in the dbEST data base (92.5% and 89.7%, respectively, in the HAP-bound fractions of methods 2-1 and 4).

With the ultimate goal of facilitating the ongoing process of gene discovery by large-scale sequencing of cDNA clones picked randomly from libraries, we have performed a pilot subtractive hybridization experiment to eliminate (or reduce representation of) a pool of approximately 5000 IMAGE Consortium-arrayed cDNA clones (pool no. 1, LLAM 78-90) from the normalized library from which they were derived (1NFLS in Table 1). PCR-amplified cDNA inserts from pool no. 1 were melted and hybridized, in the presence of blocking oligonucleotides, with single-stranded plasmid DNA from the 1NFLS library, prepared in vitro. The remaining single-stranded circles were purified by HAP chromatography, converted to double-stranded plasmids, electroporated into bacteria, and propagated under antibiotic selection to generate the subtracted 1NFLS-S1 library (see Fig. 4). Preliminary characterization of the 1NFLS-S1 library by Southern hybridization with 10 cDNA probes (only five are shown; see Fig. 5) known to be represented in pool no. 1 indicated clearly the effectiveness of the procedure to eliminate (or to reduce the representation of) all 11 cDNA sequences in the 1NFLS library. A BLASTN search of the dbEST division of GenBank (6/12/96) with 3′ ESTs obtained from the five probes (cDNAs -1, -4, -8, -9, and -10) the hybridizations of which were not shown in Figure 5, revealed the presence of 0, 0, 1, 2, and 2 corresponding ESTs, respectively, from the 1NFLS library, thus indicating that the subtraction was successful even for cDNAs that were underrepresented in the normalized library (a total of 44,407 3′ ESTs have been derived from the 1NFLS library to date). It should be noted that because of sequencing failures, some of the clones in these arrays may not yet have corresponding ESTs in the public data bases.

It is noteworthy that when we attempted to perform the same subtractive hybridization experiment using, as driver, RNA synthesized in vitro from a plasmid DNA preparation of pool no. 1, the results obtained were not satisfactory (not shown) in that subtraction could be demonstrated for some but not all tested clones (e.g, α-globin could not be subtracted effectively), similar to what we observed in normalizations with method 2-1.

DISCUSSION

As a result of an effort to improve the representation of the longest cDNAs in our normalized libraries, we have developed four different methods for normalization of directionally cloned cDNA libraries constructed in phagemid vectors, while contributing resources to the IMAGE Consortium (Lennon et al. 1996) and thereby facilitating the ongoing gene discovery and mapping programs. Approximately 87.5% of all (human) IMAGE ESTs were derived from the normalized libraries described here.

The normalization procedure (method 1) that we described previously (Soares et al. 1994) was applied for the construction of the 1NIB and 1NFLS normalized libraries, from which a total of 45,192 and 86,088 ESTs, respectively, have been derived (dbEST release 052396; http://

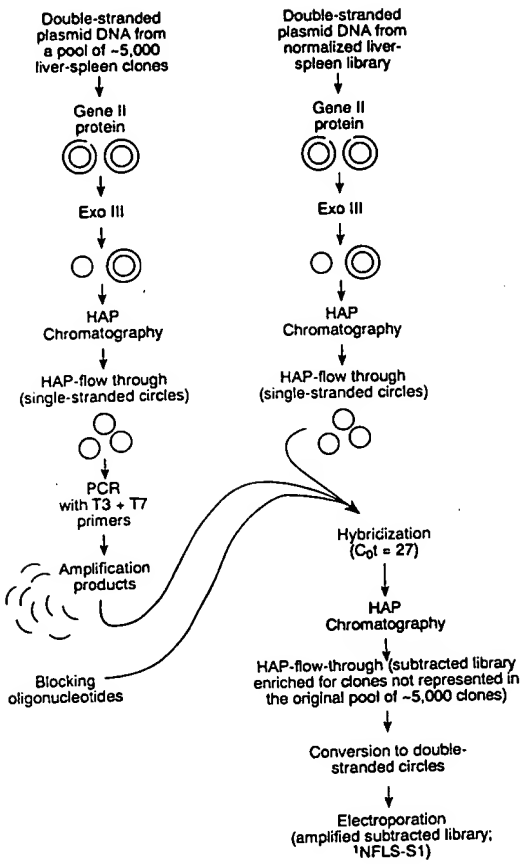

Figure 4 Diagram of the subtractive hybridization procedure used to generate the 1NFLS-S1 library. Double-stranded plasmid DNA from a pool of ~5000 IMAGE Consortium-arrayed cDNA clones derived from the 1NFLS library (pool no. 1, LLAM 78-90) was converted to single-stranded circles in vitro by the combined action of Gene II and Exonuclease III (Life Technologies). The resulting single-stranded plasmids were HAP-purified and used as a template for PCR amplification with T3 and T7 primers. PCR-amplified cDNA inserts were purified from excess primers, melted, and hybridized with single-stranded circles (prepared in vitro) from the 1NFLS library, in the presence of appropriate blocking oligonucleotides. The remaining single-stranded circles were purified by HAP chromatography, converted to double-stranded plasmids, electroporated into DH10B bacteria (Life Technologies), and propagated under ampicillin selection to generate the (1NFLS-S1) subtracted library.

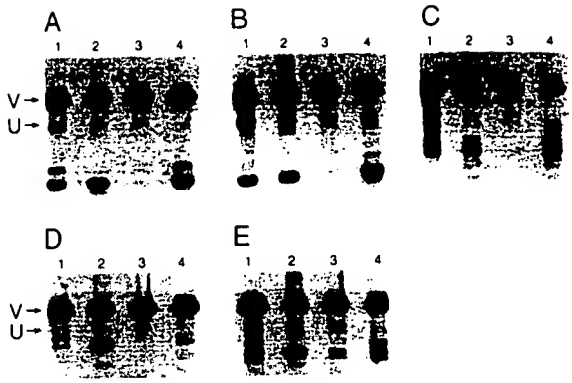

Figure 5 Characterization of the 1NFLS-S1 subtracted liver/spleen library by Southern hybridization with 5 cDNA probes. The 0.15 μg *Pac*I + *Eco*RI-digested plasmid DNA from the fetal liver/spleen library normalized with method 1 (1NFLS; lane *1*), from the pool of ~5000 IMAGE Consortium-arrayed cDNA clones derived from the 1NFLS library (pool no. 1, LLAM 78-90; lane *2*), from the subtracted library generated according to the diagram shown in Fig. 4 (1NFLS-S1; lane *3*), and from the HAP-bound fraction obtained during HAP purification of the 1NFLS-S1 library (see Fig. 4) were electrophoresed, transferred to nylon membranes, and hybridized as described in the legend to Fig. 1. The following cDNA probes were used: α-globin (*A*), γ-globin (*B*), serum albumin (*C*), unknown cDNA 7 (*D*; picked randomly from pool no. 1, LLAM 78-90), and unknown cDNA 5 (*E*; picked randomly from pool no. 1, LLAM 78-90). A BLASTN search of the dbEST subdivision of Genbank with 3′ ESTs derived from cDNA 7 and cDNA 5 revealed the presence of 33 and 0 corresponding ESTs, respectively, from the 1NFLS library. All probes were contaminated intentionally with a small amount of vector DNA to enable visualization of vector bands and thus confirm that a similar amount of library DNA was loaded in all lanes. (V) vector band; (U) residual undigested plasmid.

www.ncbi.nlm.nih.gov). Data analysis (see Hillier et al., this issue) demonstrated solidly the efficacy of this approach in bringing the frequency of all clones to within a narrow range. Extensive characterization of these two libraries by Southern analysis, however, revealed that on occasion truncated clones were favored over their longest counterparts during the normalization procedure.

Because of the relatively permissive conditions used for synthesis of first-strand cDNA, priming with the *Not*I–tag–$(dT)_{18}$ oligonucleotide may occur not only at the poly(A) tail of the mRNAs but also at internal A-rich sites within the mRNAs (e.g., at Alu tails). Typically, cDNAs with 3′ truncations occur at frequencies of 10–15% in directionally cloned libraries. Truncated clones can be recognized (tentatively) as such, by the absence of a bona fide polyadenylation signal

sequence at the appropriate distance upstream from the oligo$(dA)_{18}$ tail of the cDNA.

Why may truncated cDNAs be favored over their longest counterparts during normalization by method 1? Briefly, method 1 (Soares et al. 1994) involves: (1) annealing of a single-stranded DNA preparation of a directionally cloned cDNA library with an oligo$(dT)_{18}$ primer; (2) controlled primer extension reactions in the presence of deoxynucleotides and dideoxynucleotides to generate 3′ noncoding extension products of approximately 200–300 nucleotides; (3) purification of the resulting partially double-stranded circles by HAP chromatography; (4) melting and reassociation of the HAP-purified partially double-stranded circles to a relatively low C_0t (5–10); (5) purification of the remaining single-stranded circles (normalized library) over HAP; (6) conversion of the single-stranded circles to double-stranded circles; and (7) electroporation into bacteria.

It could be anticipated that during the reassociation reaction, because truncated cDNAs occur at lower frequencies than their nontruncated counterparts, the extension products of the truncated cDNAs would more likely reanneal to the nontruncated overlapping cDNAs than to their own truncated templates. On the other hand, the extension products of the nontruncated cDNAs would most likely reassociate to their own nontruncated templates not only because they are more prevalent but also because of the low probability of there being an overlap between the short extension product of a nontruncated clone and a truncated single-stranded circle. As a result, nontruncated single-stranded circles are more likely to end up reassociated with more than one (nonoverlapping) extension product, whereas their truncated counterparts would remain single-stranded and therefore end up in the HAP flow-through fraction (normalized library).

Reasoning that this problem could be circumvented if the hybridizing fragments (1) were in excess over single-stranded circles, and (2) spanned the entire length of the cDNAs to maximize the opportunity of overlap between truncated and nontruncated clones, we devised an approach (methods 2-1 and 2-2; note that 2-2 is the same as 2-1 except that hybridization conditions were different) whereby in vitro synthesized RNA from a plasmid DNA preparation of a starting library is used as driver in hybridization ($C_0t \sim 5$) with the same library in the form of single-stranded circles. Indeed, these modifications improved successfully the representation of the longest cDNAs in the normalized libraries (e.g., serum albumin in the liver/spleen libraries).

However, in every library constructed with methods 2-1 and 2-2, we were able to identify cDNA clones that seemed to become normalized with much greater difficulty than others (e.g., α-globin in the 5Nb2HFLS20W liver/spleen library, and G3PD in the breast library). We interpreted these results as suggestive that not all clones might be transcribed in vitro with the same efficiency if in a mixture (i.e., in vitro transcription of plasmid DNA from an entire library), and/or secondary structures in the RNAs (or interactions between RNAs) might impair their ability to hybridize with the single-stranded circles. These hypotheses were corroborated by the observation (not shown) that relatively weak hybridization signals were observed when Northern blots of RNA transcribed in vitro from an entire plasmid library were hybridized with cDNA probes derived from those clones that could not be normalized as effectively, despite the fact that they occurred at high frequencies in the starting libraries from which the in vitro transcribed RNAs were synthesized. We did exclude the possibility that the clones that were not being normalized effectively carried deletions that prevented them from being transcribed appropriately in vitro (not shown). In fact, all clones that were tested individually for in vitro transcription yielded the expected amounts of full-length RNA. Although this problem was significantly minimized in method 2-3 (cf. lanes, 1,2 in Fig. 1A–D,F,G) the extent of normalization that was achieved was still not comparable to that obtained with method 1 (cf. lanes 2,4 in Fig. 1A–D,F,H).

The advantage of method 2-3 over methods 2-1 and 2-2 is that the RNA driver is derived from a mini-library (of relatively low complexity) enriched for abundant cDNAs rather than from the entire starting library. For this reason, higher C_0t hybridizations can be carried out to eliminate or reduce significantly the representation of the most abundant cDNAs. It should be noted, however, that method 2-3 is not a true normalization procedure, because the aim of this approach is not to equalize the frequency of all cDNA clones but rather to reduce significantly (or even to eliminate, depending on the C_0t used) the representation of the most abundant clones.

The extent to which the enrichment for abundant transcripts can be achieved in such

mini-libraries depends essentially on the C_0t used for reassociation. Calculations based on estimates of frequencies of brain mRNAs (Soares et al. 1994) indicate that the best enrichments are obtained at a $C_0t = 5–10$. If the C_0t is too low (≤ 1) the enrichment is only for the most prevalent (class I) mRNAs; there is no enrichment for the mRNAs of the intermediate frequency class (class II) mRNAs. On the other hand, if the C_0t is too high (≥ 50) the enrichment for class I transcripts starts to become less significant because of a higher representation of mRNAs of the complex class (class III). Prevalent and intermediate (classes I + II) brain mRNAs comprise 93–95% of the total cDNA population in a $C_0t = 5–10$ HAP-bound mini-library, in contrast to 62% in the starting library. Consequently, the frequency of class III transcripts in a $C_0t = 5–10$ HAP-bound mini-library is about 5.5-fold lower than that of the starting library (5–7% in the bound mini-library vs. 38% in the starting library).

Methods 3 and 4 were developed as a result of an attempt to achieve both the adequate extent of normalization obtained with method 1 and the improved representation of the longest cDNAs accomplished with methods 2-1, 2-2, and 2-3. Although more technically cumbersome, method 3 is superior to method 4 in that the DNA driver used in the hybridization is single-stranded.

Single-stranded driver in method 3 (see Fig. 3) is generated by Exonuclease III digestion of gel-purified double-stranded cDNA inserts excised from the starting library. The resulting non-complementary single-stranded fragments represent the 5′ and 3′ halves of the original cDNA inserts. The fragments that correspond to the 5′ halves of the cDNAs are complementary to single-stranded circles prepared in vitro, whereas the single-stranded fragments that correspond to the 3′ halves of the cDNA inserts are complementary to single-stranded plasmids prepared in vivo. Note that for the multiple sclerosis plaques library constructed with method 3 we used single-stranded circles prepared in vitro.

Production of single-stranded circles in vitro by the combined action of Gene II and Exonuclease III (Life Technologies), rather than in vivo by superinfection of a culture with a helper phage, is very beneficial because it circumvents the distortions that otherwise may arise as a result of the differential growth properties of clones with different size inserts. However, because the digestion with Gene II results in the conversion of most, but not all, supercoiled plasmids to relaxed circles, it becomes necessary to purify the single-stranded circles that are produced after digestion with Exonuclease III by HAP chromatography.

For construction of the normalized multiple sclerosis plaques library, the cDNA inserts were excised by double digestion of plasmid DNA from the starting library with *Not*I and *Eco*RI. The fact that one in every three clones might have an internal *Eco*RI site (an *Eco* RI site is expected to occur once every 4096 bp, and the average insert size in these libraries is of the order of 1.4 kb) should not compromise the efficiency of the procedure, because at least one of the resulting restriction fragments would be expected to be ≥ 200 bp (clones smaller than 400 bp are size-selected out of these libraries) and therefore be able to form hybrids that would bind quantitatively to HAP under our conditions. A disadvantage of method 3, as presented, is that only clones <2.9 kb (approximate vector size) can be excised cleanly from the vector. It is conceivable, however, that one might be able to use double-stranded cDNA fragments generated by PCR amplification with T3 and T7 primers as substrate for the Exonuclease III digestion in method 3.

Method 4 was used to generate a significant fraction of the libraries that were contributed to the IMAGE Consortium (see Table 1). It is undoubtedly the simplest and overall most advantageous of all procedures. Because the DNA driver is generated by PCR amplification of the starting (double-stranded or single-stranded, see below) plasmid library with T3 and T7 primers, the tracer (single-stranded circles) used in this hybridization may be produced in vitro or in vivo.

The extent of normalization achieved with method 4 was comparable to that obtained with method 1 with the advantage that it successfully preserved the representation of the longest cDNAs (cf. lanes 4,5 in Fig. 1). Moreover, method 4 is superior to method 1 because it does not preclude the clones derived from mRNAs with internal *Not*I sites from being represented in the normalized library. Because the starting material for the reassociation kinetics reaction in method 1 is generated by a controlled primer extension reaction with an oligo$(dT)_{18}$ primer, clones without an oligo$(dA)_{18}$ tail (derived from mRNAs with an internal *Not*I site) are not represented in the final normalized library, although they are not necessarily lost (clones without tails end up in the HAP flow-through fraction during HAP purification of the partially double-stranded circles

generated by this primer extension reaction). It should also be noted that this problem of method 1 could be circumvented by the use of an oligonucleotide complementary to flanking vector sequences [as opposed to the oligo$(dT)_{18}$] for this controlled primer extension reaction.

The potential biases introduced by PCR amplification in method 4 are minimized by the fact that (1) PCR amplification products are used in excess in these hybridizations, and (2) the size distribution of inserts in these libraries is relatively narrow (ranging typically from 0.4 to 2.5 kb).

The conditions used for hybridization greatly influenced the quality of the resulting normalized libraries constructed with method 4. This is to a great extent a consequence of the fact that we are using HAP to purify single-stranded circles, as opposed to a biotin-avidin capture system, which in our hands yielded significantly less satisfactory results (M.F. Bonaldo and M.B. Soares, unpubl.). The best results were obtained when the hybridization conditions were the most similar to the HAP conditions. We interpreted these results as suggestive of the fact that imperfect hybrids formed during hybridization may either not bind to HAP and/or may melt once in the HAP buffer.

It is noteworthy that a much superior extent of normalization was obtained with method 4 when single-stranded plasmid DNA prepared in vitro, as opposed to double-stranded plasmid DNA, was used as template for PCR amplification (not shown). These results suggest that a fraction of the double-stranded plasmids used as template for PCR amplification, presumably in the form of melted supercoiled DNA, might end up in the HAP flow-through fraction (normalized library) during purification.

It is noteworthy that cross-hybridizing diverged sequences seem to escape normalization in all of the procedures discussed above. For example, the frequency of Alu repeat-containing cDNAs (typically 10% in directionally cloned cDNA libraries) is practically the same in starting and normalized libraries. These results suggest that imperfect hybrids either do not bind to HAP under our conditions or melt once diluted in the (more stringent) HAP buffer. This is advantageous, not only because it preserves the representation of Alu-containing cDNAs that might correspond to otherwise rare mRNAs, but also, and most significant, because it minimizes the likelihood that a rare member of a gene family might be excluded from the final (normalized or subtracted) library as a result of a cross-hybridization with a more prevalent but diverged sequence.

The use of normalized libraries for large-scale gene discovery/EST programs is beneficial because it minimizes redundancies while increasing the representation of the rarer cDNAs by about threefold, on average. However, given the great extent of overlap in gene expression among different tissues, the use of normalized libraries alone is not sufficient to maintain a desirable pace of identification of novel sequences at advanced stages of such programs. For this reason, we propose that the use of subtracted libraries enriched for clones not yet identified might become increasingly advantageous. Toward this goal, we have developed a subtractive hybridization approach designed specifically for this purpose (see Fig. 4). In a pilot experiment, we were able to reduce significantly the representation of ~5000 1NFLS–IMAGE Consortium clones from the 1NFLS library itself (see Fig. 5). With the development of appropriate clustering algorithms, the use of nonredundant sets of cDNA/gene sequences as drivers for hybridizations to generate subtractive libraries enriched for novel sequences should soon become possible, and hopefully will facilitate the isolation of all human and mouse cDNAs still awaiting identification.

METHODS

Construction of Directionally Cloned cDNA Libraries

Poly(A)$^+$ RNA was purified from total cellular RNA (except for senescent fibroblasts from which cytoplasmic RNA was isolated) using the Oligotex mRNA kit (Qiagen) according to the manufacturer's instructions, except that two rounds of purification were performed. cDNA library construction was essentially as described before (Adams et al. 1993b; Soares 1994). Typically, 1 μg poly(A)$^+$ RNA was annealed at 37°C with a twofold mass excess of a *Not*I–tag–$(dT)_{18}$ primer [or *Pac*I–tag–$(dT)_{18}$ in the case of the liver/spleen library] and reverse transcribed at 37°C with Superscript Reverse Transcriptase (Life Technologies). Alternatively poly(A)$^+$ RNA was annealed at 45°C with a fourfold mass excess of a *Not*I–tag–$(dT)_{25}$ primer and reverse transcribed at 45°C. The tag is a sequence of 2–6 nucleotides that is unique for each library and thus serves as an identifier (see Table 1). With the exception of infant brain, fetal liver/spleen and term placenta, all other first-strand cDNA syntheses were primed with the following oligonucleotide: TGTTACCAATCTGAAGTGGGAGCGGCCGC–tag–$(dT)_{18 \text{ or } 25}$. The oligonucleotide AACTGGAAGAATTCGCGGCCGCAGGAA$(dT)_{18}$ (Pharmacia) was used to prime both infant brain and term placenta first-strand cDNA syntheses. The oligonucleotide AACTGGAAGAATTAATTAAAGATCT$(dT)_{18}$ was used to prime the synthesis of first-

strand fetal liver/spleen cDNA. Double-stranded cDNAs were size-selected by gel filtration over a long (64-cm) and narrow (0.2-cm diameter) Bio-Gel A-50m (Bio-Rad, 100–200 mesh) column, and ligated to a 500- to 1000-fold molar excess of adapters. Infant brain cDNAs were ligated to *Hin*dIII adapters, digested with *Not*I, size selected over a second Bio-Gel column, and cloned directionally into the *Not*I and *Hin*dIII sites of the Lafmid BA vector (Soares et al. 1994). Fetal liver/spleen cDNAs were ligated to *Eco*RI adapters (Pharmacia), size-selected as above, digested with *Pac*I and cloned directionally into the *Pac*I and *Eco*RI sites of the pT7T3–*Pac* vector. All other cDNAs were ligated to *Eco*RI adapters (Pharmacia), size-selected as above, digested with *Not*I and cloned directionally into the *Not*I and *Eco*RI sites of the pT7T3–*Pac* vector. pT7T3–*Pac* is essentially the same as pT7T318D (Pharmacia) with a modified polylinker. Figure 6 shows the sequence of the pT7T3–*Pac* polylinker and flanking sequences.

Production of Purified Covalently Closed Single-stranded Library DNA in Vitro

Double-stranded phagemid DNA was converted to single-stranded circles by the combined action of Gene II (phage F1 endonuclease) and *Escherichia coli* Exonuclease III enzymes, as per the manufacturer's instructions (Life Technologies; cat. no. 10356-020). The resulting single-stranded circular DNA was purified from the remaining double-stranded plasmids by HAP chromatography (Bio-Rad) as described previously (Soares et al. 1994). The replication initiator protein of bacteriophage f1 (Gene II) is a site-specific endonuclease that binds to the f1 origin in phagemid vectors and nicks the viral strand of the supercoiled DNA . The nicked strand is then digested from its 3′ end with Exonuclease III (Hoheisel 1993) to generate single-stranded circles. Purification of the resulting single-stranded circles over HAP is necessary because the conversion of supercoiled to relaxed plasmids by Gene II is never complete. The Gene II reaction was performed for 1 hr at 30°C and contained typically 4 μg supercoiled plasmid library DNA, 1 μl Gene II (Life Technologies), and 2 μl 10× Gene II buffer (Life Technologies) in a total volume of 20 μl. The Gene II protein was heat inactivated for 5 min at 65°C; the reaction mixture was chilled on ice; 2 μl Exonuclease III (Life Technologies, Cat. No. 18013-011, 65 units/μl) was added; and the reaction was incubated for 30 min at 37°C. Gene II and Exonuclease III were then digested with Proteinase K (Boehringer Mannheim) for 15 min at 50°C in a 100-μl reaction containing 10 mM Tris (pH 7.8), 5 mM ethylenediamine tetraacetic acid (EDTA), 0.5% SDS, and 136 μg Proteinase K. After extraction with equal volume of phenol–chloroform–isoamyl alcohol (25:24:1), library DNA was ethanol-precipitated and digested with *Pvu*II for 2 hr at 37°C. This was done to convert the remaining supercoiled plasmids into linear DNA molecules and thereby improve their bindability to HAP under our conditions. Note that *Pvu*II does not cleave single-stranded circles and that there are two *Pvu*II sites in the vector. The reaction was diluted with 2 ml loading buffer [0.12 M sodium phosphate buffer (pH6.8), 10 mM EDTA, and 1% SDS] and purified by HAP chromatography at 60°C, using a column pre-equilibrated with the same buffer (1-ml bed vol.; 0.4 g of HAP). After a 6-ml wash with loading buffer, this volume was combined with the flow-through fraction, and the sample was extracted twice with water-saturated 2-butanol, once with dry 2-butanol, and once with water-saturated ether (3 vols. per extraction). Residual ether was blown off by vacuum and the sample was desalted by passage through a Nensorb column (DuPont/NEN) according to the manufacturer's specifications, concentrated down to ~0.35 ml and ethanol-precipitated. Note that Gene II–Exonuclease III prepared single-stranded DNA is in the opposite polarity to single-stranded DNA generated by in vivo phagemid production.

Production of Purified Covalently Closed Single-stranded Library DNA in Vivo

Plasmid DNA from the starting library was electroporated into *E. coli* DH5αF′ bacteria, and the culture was grown under ampicillin selection at 37°C to an OD_{600} of 0.2, superinfected with a 10- to 20-fold excess of the helper phage M13KO7 (Pharmacia), and harvested after 4 hr for preparation of single-stranded plasmids, as described (Vieira and Messing 1987).

Conversion of Single-stranded Circles to Double-stranded Plasmids

Single-stranded circles (<50 ng) were ethanol-precipitated and resuspended in 11 μl water. Then 4 μl 5× Sequenase buffer (USB) and 1 μl primer (1 μg) were added and the mixture was incubated at 65°C for 5 min and then at 37°C for 3 min. Then 1 μl Sequenase version 2.0 (USB), 1 μl 0.1 M dithiothreitol (DTT), and 2 μl mixed dNTP stock (a solution containing each deoxynucleotide at a final concentration of 10 mM) were added, and the reaction was incubated at 37°C for 30 min. The total volume was taken up to 100 μl with 10 mM Tris (pH8.0) and 1 mM EDTA (TE) and the reaction was extracted once with phenol–chloroform–isoamyl alcohol (25:24:1). Plasmid DNA was ethanol-precipitated and dissolved in 3 μl TE. The following oligonucleotides were used for this primer extension reaction: (1) M13 Reverse Sequencing Primer (5′-AGCGGATAACAATTTCACACAGGA-3′), which is complementary to single-stranded prepared in vitro,

5′-caccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcacacaggaaacagctatg
M13 Reverse Sequencing Primer

acatgattacgaatttaatacgactcactatagggaatttGGCCCTCGAGGCCAAGAATTCCCGACTACGTA
T7 Promoter *Sfi*I *Eco*RI *Sna*BI

GTCGGGGATCCGTCTTAATTAAGCGGCCGCAAGCTTattccctttagtgagggttaattttagcttggcac
*Bam*HI *Pac*I *Not*I *Hin*dIII T3 Promoter

tggccgtcgttttacaacgtcgtgactgggaaaaccctggcgttacccaacttaatcgccttgcag-3′.
M13 Sequencing Primer

Figure 6 Sequence of the pT7T3–*Pac* polylinker (uppercase) and flanking sequences (lowercase).

and (2) Oligo-Amp (5'-GACTGGTGAGTACTCAAC-CAAGTC-3'), which is complementary to the ampicillin resistance gene of single-stranded pT7T3–Pac or Lafmid BA plasmids prepared in vivo.

In Vitro Synthesis of Library RNA

Some 2–5 μg of double-stranded plasmid DNA from either the starting library (see methods 2-1 and 2-2 below) or the mini-library of abundant cDNAs (see method 2-3 below) was linearized with either *Pac*I (NEB) or *Not*I (NEB) and used as a template for synthesis of RNA with RiboMax Large Scale RNA Production Systems T7 or T3 (Promega), according to the manufacturer's instructions. After treatment with ribonuclease-free DNAse I (Promega), to digest away the plasmid DNA template, the RNA was used for hybridization as described below. It should be noted that RNA synthesized with T7 RNA Polymerase is in the message-like orientation and is complementary to the single-stranded circles produced in vitro. On the other hand, RNA synthesized with T3 RNA Polymerase is in the antimessage orientation and it is complementary to single-stranded circles produced in vivo.

Normalization Method 1

The procedure used for construction of the normalized human infant brain (1NIB) library (here designated as method 1) has been described previously (Soares et al. 1994). Method 1, with minor modifications, was also applied to construct the normalized human fetal liver/spleen cDNA library (1NFLS). To synthesize a partial second strand of about 200 nt by limited extension, a 100 μl reaction mixture containing 5 μl 0.5 μg/μl *Pvu*II-digested, HAP- and gel-purified single-stranded plasmid DNA from the fetal liver/spleen starting library, 7 μl 10 ng/μl oligo $(dT)_{12-18}$ (Pharmacia), 10 μl 10× Primer Extension Buffer [0.3 M Tris (pH 7.5), 0.5 M NaCl, and 0.15 M $MgCl_2$], 10 μl 0.1 M DTT, 10 μl mixed dNTP stock, 25 μl mixed ddNTP stock (a solution containing each dideoxy A, C, and G at a final concentration of 25 mM), 5 μl 800 Ci/mmole [α-^{32}P]dCTP, and 20.5 μl water was incubated at 60°C for 5 min, at 50°C for 15 min, and at 37°C for 2 min. Then 7.5 μl 5 units/μl Klenow enzyme (USB) was added, and the reaction was incubated at 37°C for 30 min. The reaction was extracted with phenol–chloroform–isoamyl alcohol (25:24:1), 5 μg melted and sheared salmon sperm DNA was added, and the partially double-stranded plasmids were purified from the remaining single-stranded circles (unprimed molecules, as well as clones derived from mRNAs with an internal *Pac*I site that therefore do not contain an oligo(dA) tail at the 3' end) by HAP chromatography. The HAP-bound fraction containing the partially double-stranded plasmids was eluted with 6 ml 0.4 M sodium phosphate buffer (pH 6.8), 10 m M EDTA, and 1% SDS, and plasmid DNA was desalted as described before (Soares et al. 1994) and ethanol-precipitated. The DNA (173 ng) was resuspended in 2.5 μl deionized formamide and melted at 80°C for 3 min under 10 μl mineral oil. Then 1 μl of 5 μg/μl oligo $(dT)_{12-18}$ (used to block the tails) was added, and the mixture was heated at 80°C for 1 min. Then 0.5 μl 5 M NaCl, 0.5 μl 10× TE, and 0.5 μl water were added, and the reassociation reaction was incubated at 42°C for 0.6 hr (calculated $C_0t = 0.5$). The remaining single-stranded circles were purified over HAP (flow-through fraction) and subjected subsequently to a second cycle of the normalization procedure as described above, except that reassociation was conducted for 24 hr (calculated $C_0t = 20$). The remaining single-stranded circles (normalized library; 1NFLS) were purified over HAP, converted to double-stranded plasmids, electroporated into DH10B bacteria, and propagated under ampicillin selection.

Normalization Methods 2-1, 2-2, and 2-3

Method 2 is a reassociation kinetics-based approach involving hybridization of in vitro synthesized RNA (the driver) derived either from the entire library (methods 2-1 and 2-2; see Fig. 2) or from a mini-library enriched for abundant cDNAs (method 2-3; see Fig. 2), with the whole starting library in the form of single-stranded circles (the tracer). The remaining single-stranded circles (normalized library) are purified by HAP chromatography (HAP flow-through fraction), converted to double-stranded plasmids for improvement of electroporation efficiency, electroporated into DH10B bacteria (Life Technologies), and propagated under ampicillin selection. A number of normalized cDNA libraries were constructed with these methods using single-stranded plasmids prepared both in vivo and in vitro (see Table 1). In all three variants, the driver was first pre-annealed with a pair of oligonucleotides to block both 5' and 3' vector sequences as follows: 0.5 μl (10 μg) of each oligonucleotide, 1 μl RNA (5.0 μg in methods 2-1 and 2-3; 0.5 μg in method 2-2), and 4.0 μl deionized formamide were heated for 3 min at 80°C under 10 μl mineral oil and quickly chilled on ice. Then 0.8 μl 10× hybridization buffer [0.4 M Pipes (pH 6.4), 4 M NaCl, and 10 mM EDTA in methods 2-1 and 2-3; 0.4 M Pipes (pH 6.4), 1.2 M NaCl, 10 mM EDTA, and 1% SDS in method 2-2), 0.5 μl RNAsin (Boehringer Mannnheim), and 0.7 μl water were added and the mixture (total volume, 8 μl) was incubated overnight at 42°C (methods 2-1 and 2-3) or 30°C (method 2-2). In another tube, 2.5 μl (50 ng) single-stranded library DNA in deionized formamide was heated for 3 min at 80°C under mineral oil; 0.5 μl 10× hybridization buffer and 2.0 μl water were added; and the mixture was transferred to the tube containing the preannealed RNA. Hybridization (13-μl reaction) was performed at 42°C (method 2-1: C_0t = 5–10; method 2-3: C_0t = 100–200) or at 30°C (method 2-2: C_0t = 5–10). The driver, rather than the tracer, was blocked because otherwise the latter would, to some extent, bind to HAP during purification. The plasmid mini-library enriched for abundant cDNAs that served as a template for the synthesis of RNA used as driver in method 2-3 was prepared from the HAP-bound fraction obtained during purification of the normalized library in method 2-1. Different pairs of blocking oligonucleotides were used, depending on whether the RNA was synthesized with T3 or T7 RNA polymerases. To block RNA synthesized with T3 RNA polymerase, which was used in hybridizations with single-stranded plasmids prepared in vivo we used: 5'-$_{19}$AGGGCGGCCGCAAGCTTATTCCCTTTAGT-GAGGGTTAAT-3' (this oligonucleotide was used to block 5' vector sequences of all but the human fetal liver/spleen library RNA), and 5'-$_{19}$AGATCTTTAATTAAGCGGCCG-CAAGCTTATTCCCTTTAGTGAGGGTTAAT-3' (this oligonucleotide was used to block 5' vector sequences of the

human fetal liver/spleen library RNA), and 5′-AGGCCAAGAATTCGGCACGAG-3′ (this oligonucleotide was used to block 3′ vector sequences). To block RNA synthesized with T7 RNA polymerase, which was used in hybridizations with single-stranded plasmids prepared in vitro we used: 5′-CCTCGTGCCGAATTCTTGGCCTCGAGGGCCAAATTCCC-3′ (this oligonucleotide was used to block 5′ vector sequences). The oligonucleotide used to prime the synthesis of first-strand cDNA was also used to block 3′ vector sequences.

Normalization Method 3

Method 3, used to generate the normalized library from multiple sclerosis plaques (2NbHMSP), is a reassociation-kinetics-based approach involving hybridization (C_0t = 20–25) of a 20-fold excess of Exonuclease III-digested cDNA inserts excised from a plasmid DNA preparation of the starting library with the library itself in the form of single-stranded circles, followed by HAP-purification of the remaining single-stranded plasmids, conversion to double-strands, and electroporation into bacteria. Some 5 μg double-stranded plasmid DNA from the starting library was doubly digested with *Not*I and *Eco*RI; the excised cDNA inserts were separated from the cloning vector by agarose gel electrophoresis; and the DNA was purified using beta-agarase (NEB) according to the manufacturer's instructions. Then 0.6 μg gel-purified double-stranded cDNA inserts in 47.5 μl TE was digested with Exonuclease III at 37°C for 30 min in a 60-μl reaction containing 6 μl 10× Exonuclease III buffer [0.5 M Tris (pH 8.0) and 50 mM $MgCl_2$)], 0.6 μl 0.1 M DTT, 2.9 μl water, and 3 μl of 65 units/μl Exonuclease III (Life Technologies). The Exonuclease was then digested with 136 μg Proteinase K (Boehringer Mannheim) at 50°C for 15 min in a 100-μl reaction containing 10 mM Tris (pH 7.8), 5 mM EDTA, and 0.5% SDS. After two extractions with phenol–chloroform–isoamyl alcohol (25:24:1), the resulting noncomplementary single-stranded DNA (total amount ~0.3 μg) was ethanol-precipitated and resuspended in 1 μl TE. A 5-μl hybridization reaction was then set up as follows: 1 μl Exonuclease III-digested cDNA inserts (an estimated amount of 150 ng of single-stranded DNA) and 50 ng single-stranded plasmid DNA from the starting multiple sclerosis plaques library (prepared in vitro) in 2.5 μl deionized formamide were mixed and heated at 80°C for 3 min under 10 μl mineral oil. Then 0.5 μl (10 μg) of a blocking oligonucleotide (5′-CCTCGTGCCGAATTCTTGGCCTCGAGGGCCAAATTCCCTATAGTGAGTCGTATTA-3′), 0.5 μl 5 M NaCl, and 0.5 μl 10× TE were added, and the mixture was incubated at 42°C for 41 hr (calculated C_0t of 23). The remaining single-stranded plasmids were purified by HAP chromatography, converted to double-stranded plasmids, and electroporated into DH10B bacteria (Life Technologies) as described above.

Normalization Method 4

This is a reassociation-kinetics-based approach involving hybridization of a 20-fold excess of cDNA inserts generated by PCR with the library itself in the form of single-stranded circles, followed by HAP purification of the remaining single-stranded plasmids, conversion to double-strands, electroporation into DH10B bacteria, and amplification under ampicillin selection. PCR amplification of cDNA inserts was performed using the Expand High Fidelity PCR System (Boehringer Mannheim) according to the manufacturer's instructions. This PCR system is composed of an enzyme mixture containing thermostable Taq DNA and Pwo DNA polymerases (Barnes 1994). An amount of 1 μl (2.5–5.0 ng) DNA template [double-stranded plasmids (fetal lung, parathyroid adenoma, senescent fibroblasts) or single-stranded circles prepared in vitro (fetal heart, 14Nb2HFLS20W-fetal liver/spleen, and all mouse, rat, and schistosome libraries listed in Table 1)] was mixed with 2 μl dNTP stock (the final concentration of each dNTP in the reaction is 200 μM), 5 μl of a 20-μM solution of T7 Primer (5′-TAATACGACTCACTATAGGG-3′), 5 μl of a 20-μM solution of T3 Primer (5′-ATTAACCCTCACTAAAGGGA-3′), 10 μl 10× I Expand High Fidelity buffer, 0.75 μl Expand High Fidelity enzyme mix (2.6 units), and 76.25 μl water. Then 50 μl mineral oil was added and the reaction mixture was subjected to the following amplification cycle conditions in a Perkin Elmer Thermocycler: 7 min while ramping up from room temperature to 94°C; 20 cycles of 1 min at 94°C, 2 min at 55°C, and 3 min at 72°C, and 7 min at 72°C. PCR-amplified fragments were purified using the High Pure PCR Product Purification Kit (Boehringer Mannheim) as instructed by the manufacturer. The purified PCR product was ethanol-precipitated and dissolved in 5 μl TE. Then 1.5 μl (0.5 μg) PCR products was mixed with 5 μl (50 ng) library DNA (single-stranded circles prepared in vitro) in deionized formamide, 0.5 μl (10 μg) 5′ blocking oligo AV-1 (5′-CCTCGTGCCGAATTCTTGGCCTCGAGGGCCAAATTCCCTATAGTGAGTCGTATTA-3′), 0.5 μl (10 μg) 3′ blocking oligo AR (5′-ATTAACCCTCACTAAAGGGAATAAGCTTGCGGCCGCT$_{20}$-3′; used for all but the fetal liver/spleen library), or alternatively, (0.5 μl (10 μg) 3′ blocking oligo AV-2 (5′-ATTAACCCTCACTAAAGGGAATAAGCTTGCGGCCGCTTAATTAAAGATCT$_{19}$-3′; used only for the fetal liver/spleen library), and this mixture was heated at 80°C for 3 min under 10 μl of mineral oil. Then 1 μl 10× buffer-A [1.2 M NaCl, 0.1 M Tris (pH 8.0), and 50 mM EDTA; used for fetal lung, fetal heart, parathyroid adenoma, senescent fibroblasts, and 19.5-days postconception (dpc) mouse embryo] or, alternatively, 1 μl 10× buffer-B [1.2 M NaCl, 0.1 M Tris (pH 8.0), 50 mM EDTA, and 10% SDS; used for 14Nb2HFLS20W-fetal liver/spleen, 17.5-dpc mouse embryo, 13.5- to 14.5-dpc mouse embryo, rat heart, rat kidney, and 8-week schistosome], and 1.5 μl water were added, and the hybridization was performed at 30°C for 24 hr (calculated C_0t ~ 5). The remaining single-stranded circles were purified by HAP chromatography, converted to double-strands, and electroporated into DH10B (Life Technologies) bacteria, as described above.

Subtractive Hybridization

Double-stranded plasmid DNA from a pool of 4992 clones grown individually in 384 well plates (IMAGE Consortium plates LLAM 78-90, identification nos. 66696–67079 and 108168–112775) derived from the normalized fetal liver/spleen library (1NFLS) was prepared using the Qiagen Midi-prep kit according to the manufacturer's instructions, and converted to single-stranded circles in vitro, as described above. Single-stranded circles were purified by

HAP chromatography and used as a template for PCR amplification with T7 and T3 primers, as described above. An amount of 1.5 μg of PCR-amplified cDNA inserts from the LLAM 78-90 pool (in 4 μl deionized formamide) was mixed with 50 ng of single-stranded circles from the 1NFLS library (in 2 μl deionized formamide), 2.1 μl (42 μg) 5′ blocking oligo AV-1, and 2.1 μl (42 μg) 3′ blocking oligo AV-2. Then 10 μl mineral oil was added, and the mixture was heated at 80°C for 3 min. Then 1.2 μl 10× buffer-B and 0.6 μl water were added, and the hybridization was performed at 30°C for 48 hr (calculated $C_0t = 27$). The remaining single-stranded circles were purified over HAP, converted to double-strands, electroporated into DH10B bacteria, and propagated under ampicillin selection to generate the subtracted liver/spleen library (1NFLS-S1). HAP-bound DNA was also processed and purified for use in control experiments.

ACKNOWLEDGMENTS

We are most grateful to Dr. Joel A. Jessee (Life Technologies) for helpful discussions and for having supplied us with Gene II. We are also thankful to Dr. LaDeana Hillier and Dr. Marco Marra (Genome Sequencing Center at Washington University in St. Louis) for having diligently provided us with feedback information on several features of our libraries, based on the voluminous sequence data that they obtained, which greatly facilitated our assessment of the efficacy of the various methods that we developed. We are also in debt to Dr. Stephen Brown (Columbia University), Dr. Conrad Gilliam (Columbia University), Dr. Anne Bowcock and Ms. Monique Spillman (University of Texas Southwestern Medical Center at Dallas), Dr. Donald Gilden (University of Colorado Health Sciences Center), Dr. Val Sheffield (University of Iowa), Dr. Roderick McInnes (University of Toronto and Hospital for Sick Children, Canada), Dr. David Klein [National Institute of Child Health and Human Development, National Institutes of Health (NIH)], Dr. Anthony Albino and Dr. Alice de Oliveira (Memorial Sloan-Kettering Cancer Center), Dr. Stephen Marx (National Institute of Diabetes and Digestive and Kidney Diseases, NIH), Dr. Barbara Burkhart (National Institute of Environmental Health Sciences, NIH), Dr. Kevin Becker [National Institute of Neurological Disorders and Stroke (NINDS), NIH], Dr. Minoru Ko (Wayne State University), Dr. Ronald Blanton and Dr. Aravinda Chakravarti (Case Western Reserve University), and Dr. Mark Boguski (National Centre for Biotechnology Information, NIH) for having either faciliated our access to or provided tissue or total RNA from most sources used in construction of the libraries described in this manuscript. We are also most grateful to Mr. Long Su, Dr. Pierre Jelenc, Ms. Lee Lawton, Mrs. Ling Qiu, and Ms. Susan Baumes for most valuable assistance throughout this work. This work was supported by grants from the U.S. Department of Energy (FG02-91ER61233) and the National Center for Human Genome Research, NIH (R01 HG00980), to M.B.S. The work of G.L. was performed under the auspices of the U. S. Department of Energy by Lawrence Livermore National Laboratory (LLNL) under contract number W-7405-ENG-48..

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Adams, M.D., J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno, A.R. Kerlavage, W.R. McCombie, and J. Craig Venter. 1991. Complementary DNA sequencing: Expressed sequence tags and Human Genome Project. *Science* **252:** 1651–1656.

Adams, M.D, M. Dubnick, A.R. Kerlavage, R. Moreno, J.M. Kelley, T.R. Utterback, J.W. Nagle, C. Fields, and J. Craig Venter. 1992. Sequence identification of 2,375 human brain genes. *Nature* **355:** 632–634.

Adams, M.D., A.R. Kerlavage, C. Fields, and J.C. Venter. 1993a. 3,400 new expressed sequence tags identify diversity of transcripts in humna brain. *Nature Genet.* **4:** 256–267.

Adams, M.D., M.B. Soares, A.R. Kerlavage, C. Fields, and J.C. Venter. 1993b. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nature Genet.* **4:** 373–380.

Adams, M.D., A.R. Kerlavage, R.D. Fleischmann, R.A. Fuldner, C.J. Bult, N.H. Lee, E.F. Kirkness, K.G. Weinstock, J.D. Gocayne, O. White, et al. 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature* **377:** 3–174.

Altschul, S.F., W. Gish, W. Miller, E. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.

Barnes, W.M. 1994. PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates. *Proc. Natl. Acad. Sci.* **91:** 2216–2220.

Berry, R., T.J. Stevens, N.A. Walter, A.S. Wilcox, T. Rubano, J.A. Hopkins, J. Weber, R. Goold, M.B. Soares, and J.M. Sikela. 1995. Gene-based sequence-tagged-sites (STSs) as the basis for a humna gene map. *Nature Genet.* **10:** 415–423.

Bishop, J.O., J.G. Morton, M. Rosbash, and M. Richardson. 1974. Three abundance classes in HeLa cell messenger RNA. *Nature* **250:** 199–204.

Davidson, E.H. and R.J. Britten. 1979. Regulation of gene expression: Possible role of repetitive sequences. *Science* **204:** 1052–1059.

Hillier, L., G. Lennon, M. Becker, M. Bonaldo, B. Chiapelli, S. Chissoe, N. Dietrich, T. DuBuque, A. Favello, W. Gish, et al. 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* (this issue).

Hoheisel, J.D. 1993. On the activities of Escherichia coli exonuclease III. *Anal. Biochem.* **209:** 238–246.

Houlgatte, R., R. Mariage-Samson, S. Duprat, A. Tessier, S. Bentolila, B. Lamy, and C. Auffray. 1995. The Genexpress Index: A resource for gene discovery and genic map of the human genome. *Genome Res.* **5:** 272–304.

Johnston, S., J.H. Lee, and D.S. Ray. 1985. High-level expression of M13 gene II protein from an inducible polycistronic messenger RNA. *Gene* **34:** 137–145.

Khan, A.S., A.S. Wilcox, M.H. Polymeropoulos, J.A. Hopkins, T.J. Stevens, M. Robinson, A.K. Orpana, and J.M. Sikela. 1992. Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nature Genet.* **2:** 180–185.

Lennon, G.G., C. Auffray, M. Polymeropoulos, and M.B. Soares. 1996. The I.M.A.G.E. Consortium: An integrated molecular analysis of genomes and their expression. *Genomics* **33:** 151–152.

Combie, W.R., M.D. Adams, J.M. Kelley, M.G. FitzGerald, T.R Utterback, M. Khan, M. Dubnick, A.R. Kerlavage, J.C. Venter, and C. Fields. 1992. *Caenorhabditis elegans* expressed sequence tags identify gene families and potential disease gene homologues. *Nature Genet.* **1:** 124–131.

Okubo, K., N. Hori. R. Matoba, T. Niiyama, A. Fukushima, Y. Kojima, and K. Matsubara. 1992. Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genet.* **2:** 173–179.

Rasched, I. and E Oberer. 1986. Ff coliphages: Structural and functional relationships. *Microbiol. Rev.* **50:** 401–427.

Soares, M.B. 1994. Construction of directionally cloned cDNA libraries in phagemid vectors. In *Automated DNA sequencing and anlaysis* (ed. M.D. Adams, C. Fields, and J.C. Venter), pp. 110–114. Academic Press, New York, NY.

Soares, M.B., M.F. Bonaldo, P. Jelenc, L. Su, L. Lawton, and A. Efstratiadis. 1994. Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci.* **91:** 9228–9232.

Vieira, J. and J. Messing. 1987. Production of single-stranded plasmid DNA. *Methods Enzymol.* **153:** 3–11.

Received June 14, 1996; accepted in revised form July 29, 1996.